Running LLMs Locally With Docker Model Runner

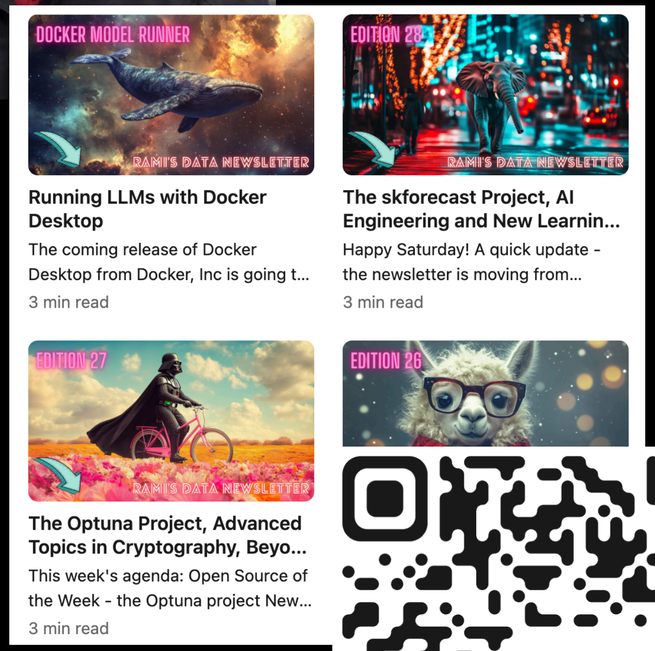
ODSC AI West 2025

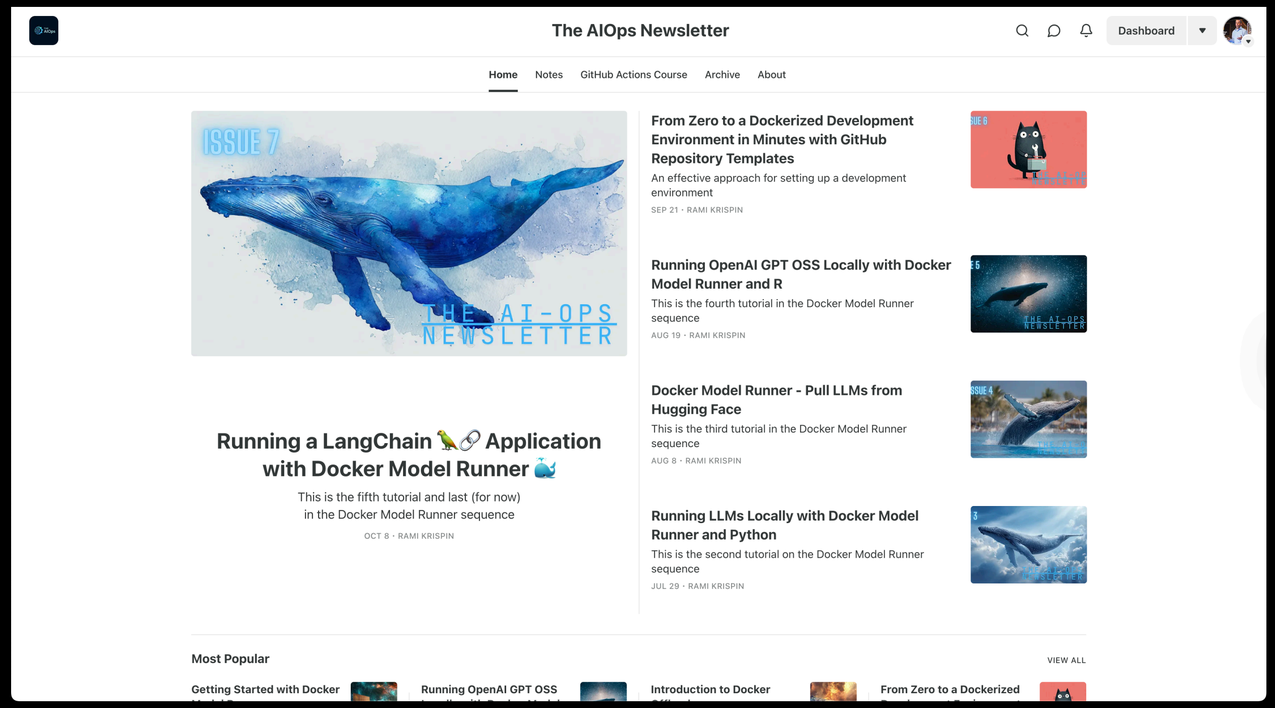
Rami Krispin, October 28th, 2025

Rami Krispin

Senior Manager - Data Science & Engineering

Author | Docker Captain 🐋 | LinkedIn Instractor





Agenda

- Introduction to DMR
- Setup
- Pull LLMs from Docker Hub & Hugging Face
- Running LLMs via the CLI
- Running LLMs with Python
- OpenCode

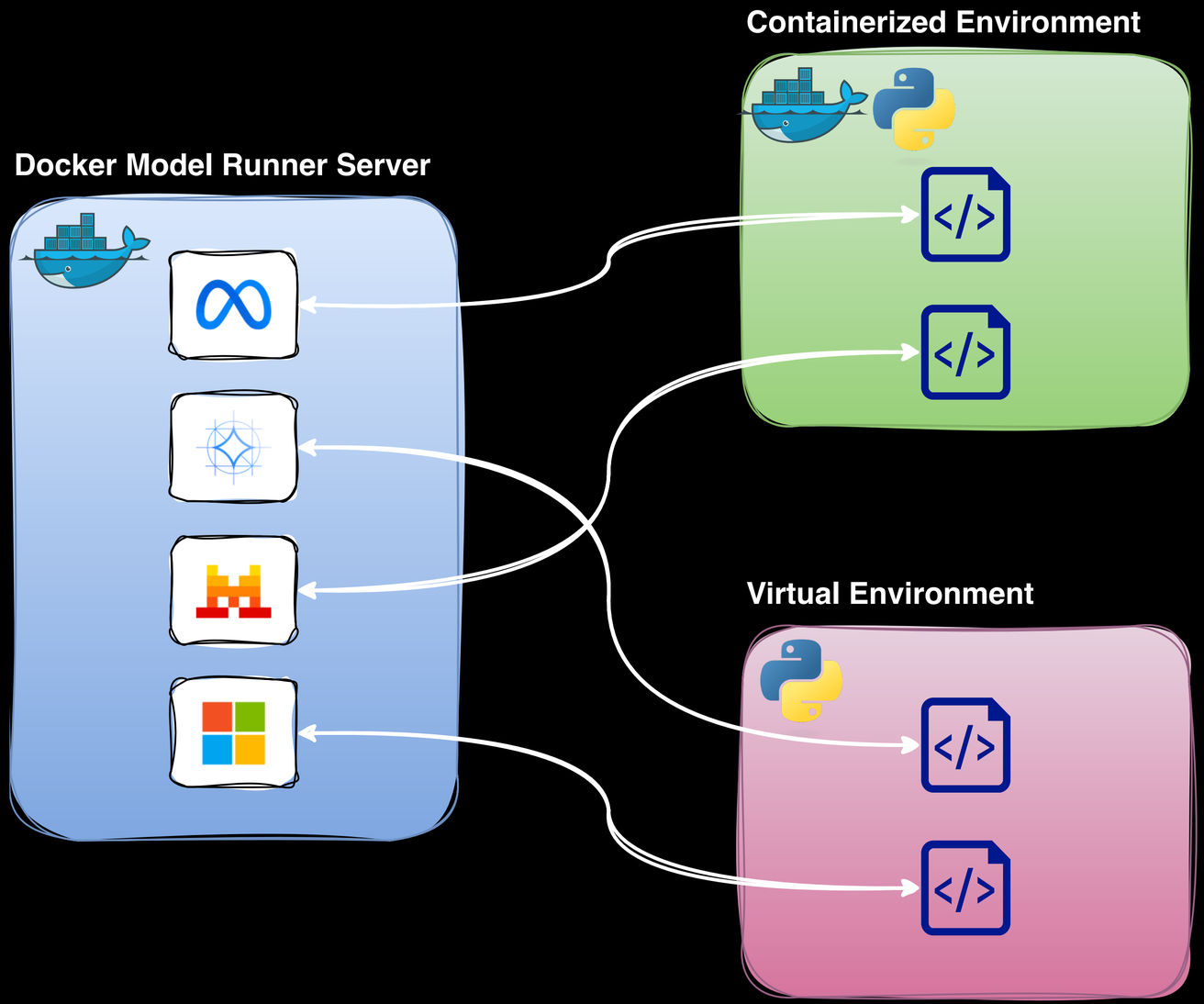
Poll

- Docker
- Python
- OpenAI Python SDK

https://github.com/RamiKrispin/odsc-ai-2025-dmr-workshop

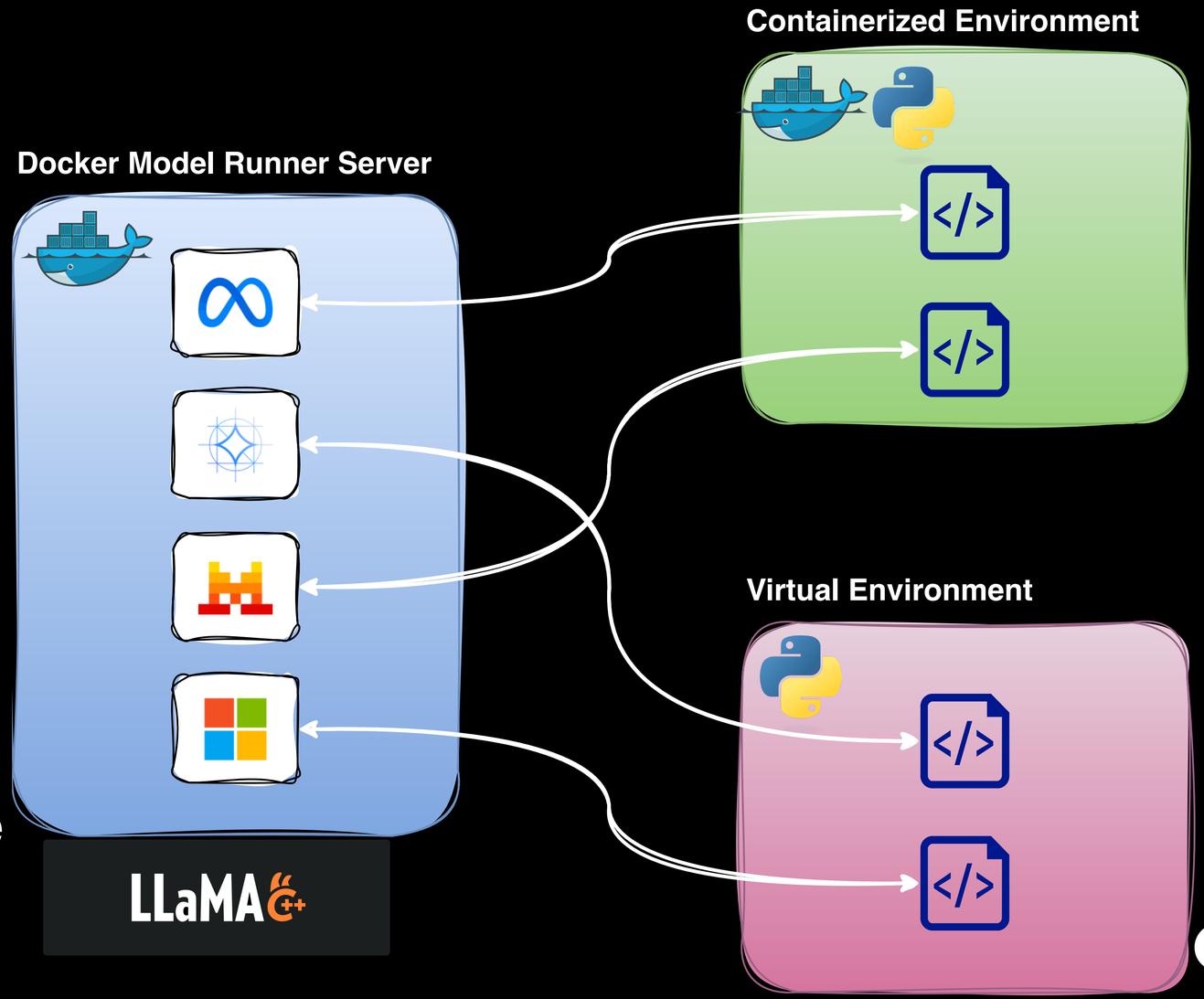
DMR Doesn’t Require Prior Docker Knowledge

Introduction to DMR



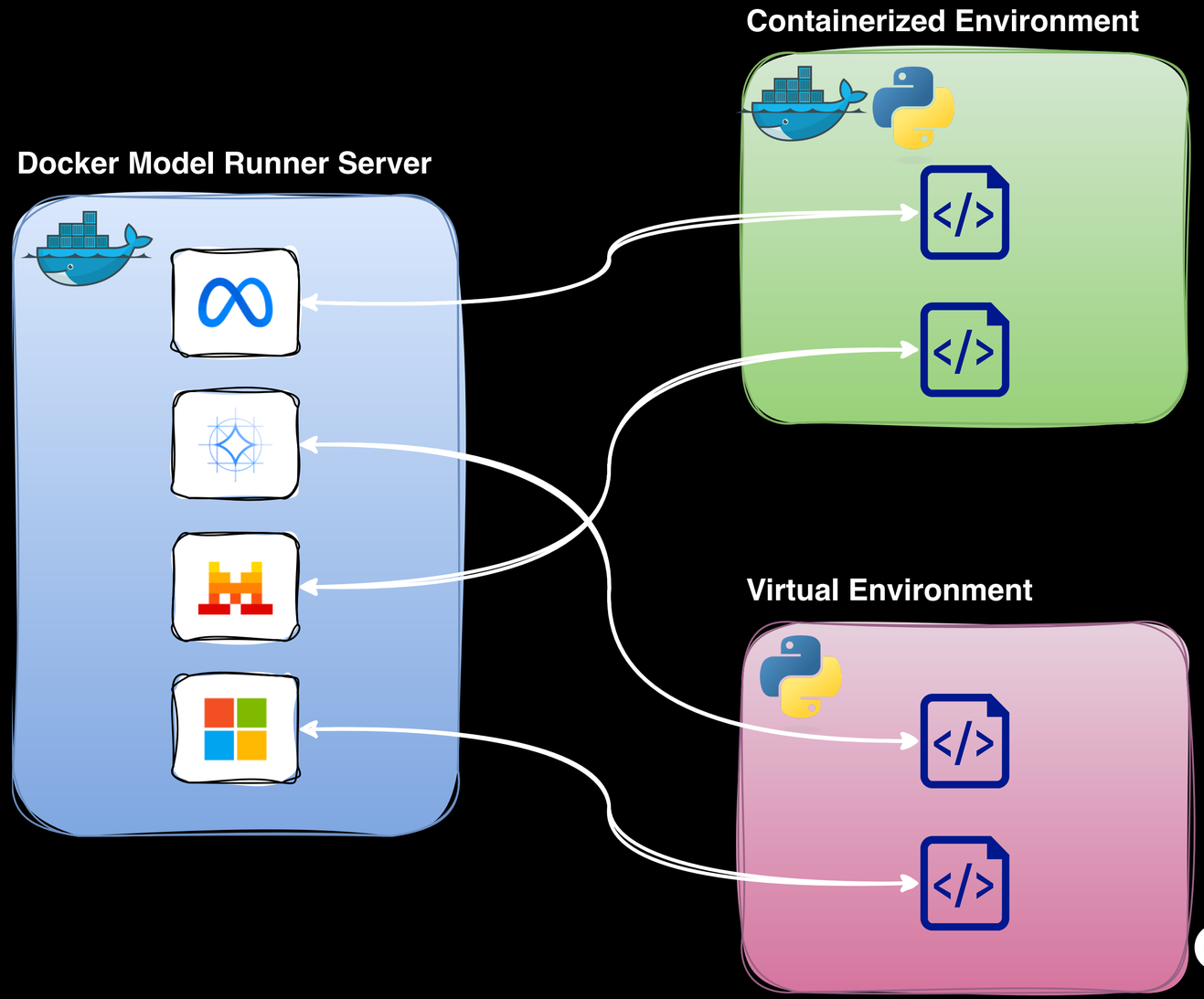
Introduction to DMR

- Docker Desktop feature
- Based on Llama.cpp
- Easy to setup and use
- Open Container Initiative
- Supports GGUF format
- OpenAI API SDK compatible



Applications

- Local development of AI applications
- Running local AI applications (local AI agents, code generation, etc.)
- Privacy



DMR Backend

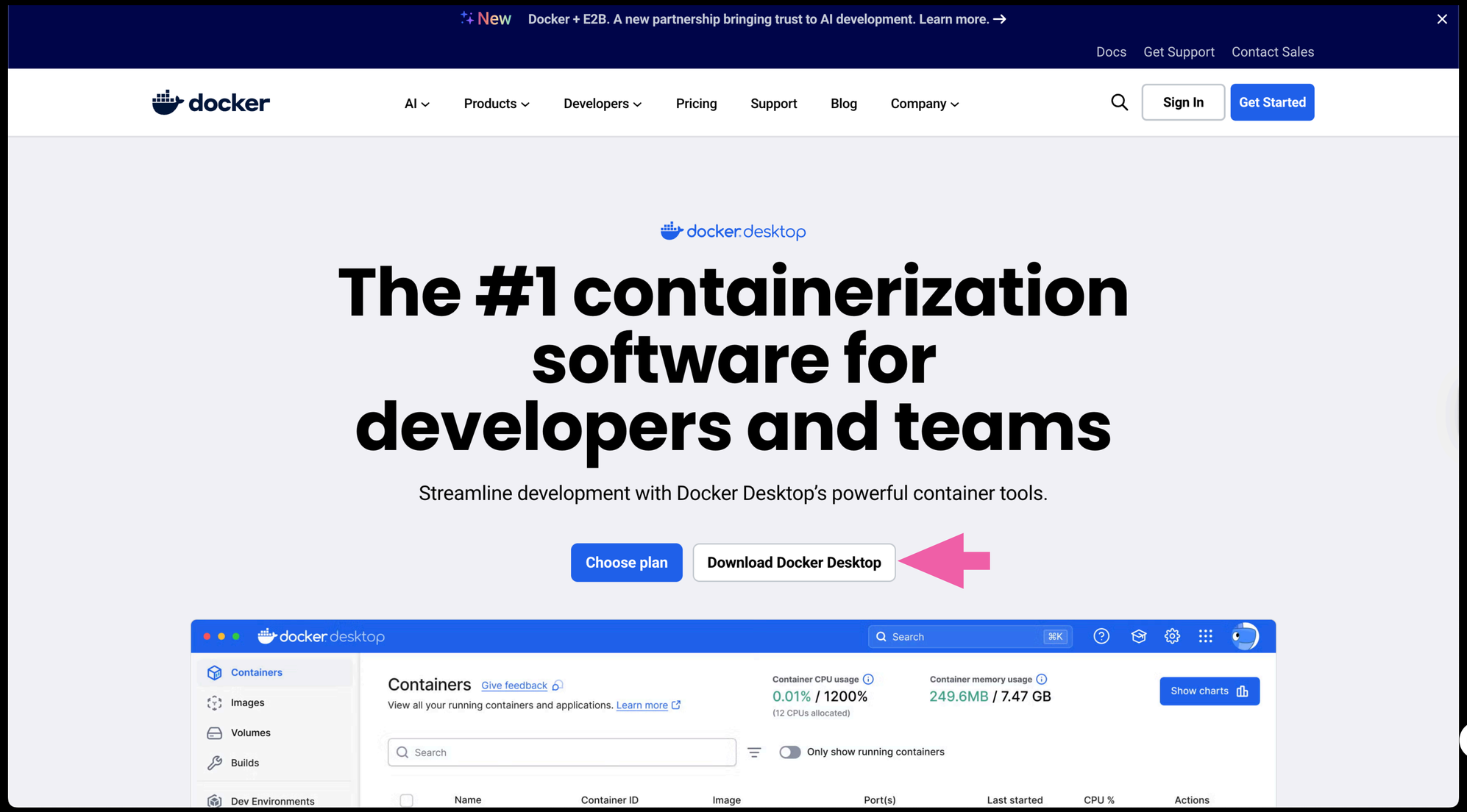

AI GUIDE TO THE GALAXY
Running
Local LLMs
with Docker
Model Runner
docker

General Requirements

- Docker Engine (Linux)
- Docker Desktop 4.40 (Mac) or 4.41 (Windows)
- Docker Hub account
- Hugging Face API key (optional)
- OpenCode (optional)
- Local computational resources

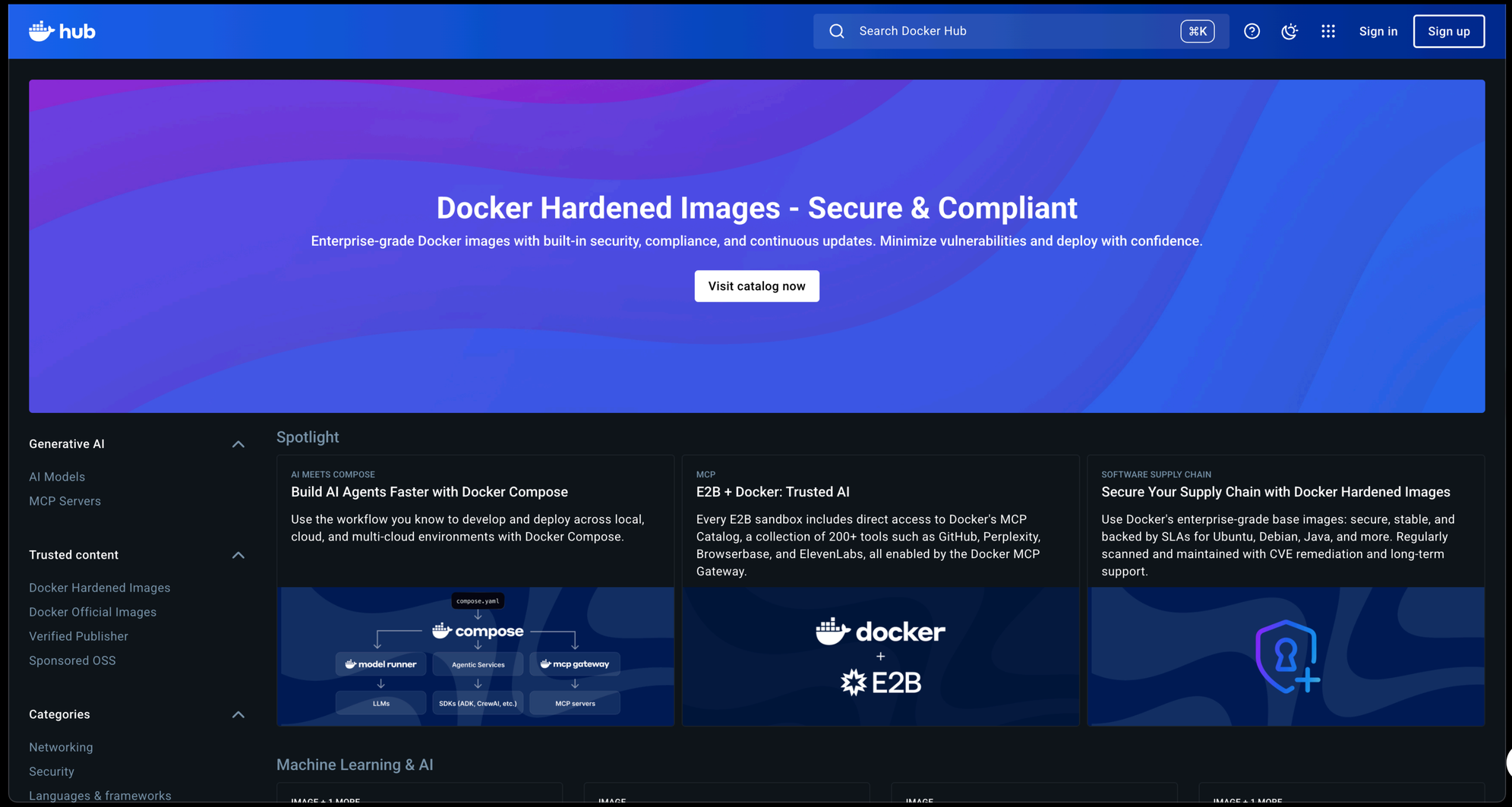
Setup

Setup

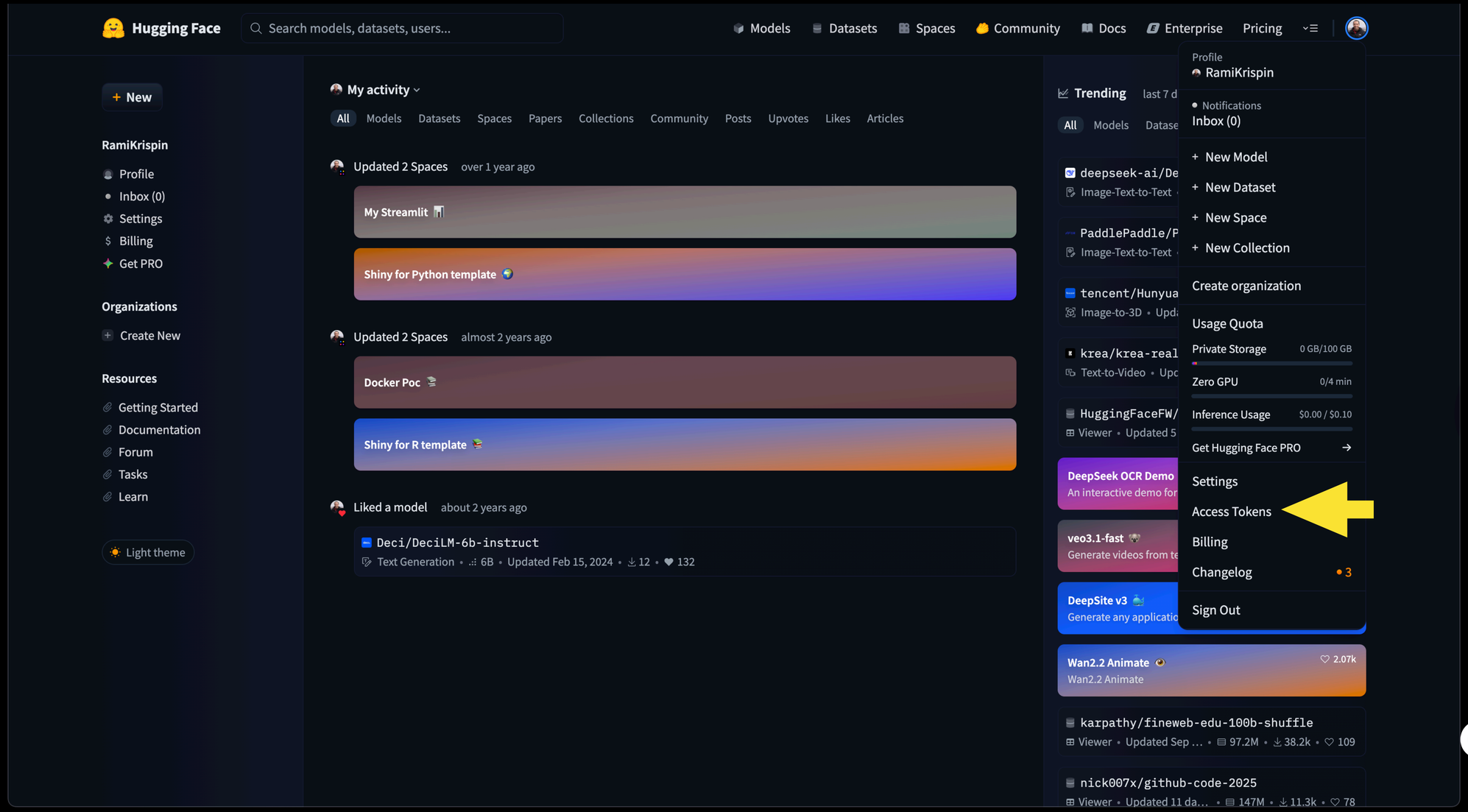


ODSC AI WEST 2025

Setup



Setup



Setup

opencode

- Intro
- Config
- Providers
- Enterprise
- Troubleshooting

Usage

- TUI
- CLI
- IDE
- Zen
- Share
- GitHub
- GitLab

Configure

- Tools
- Rules
- Agents
- Models
- Themes
- Keybinds
- Commands
- Formatters
- Permissions
- LSP Servers

Install

The easiest way to install OpenCode is through the install script.

```
curl -fsSL https://opencode.ai/install | bash
```

You can also install it with the following commands:

Using Node.js

NPM | BUN | PNPM | YARN

```
npm install -g opencode-ai
```

Using Homebrew on macOS and Linux

```
brew install sst/tap/opencode
```

Using Paru on Arch Linux

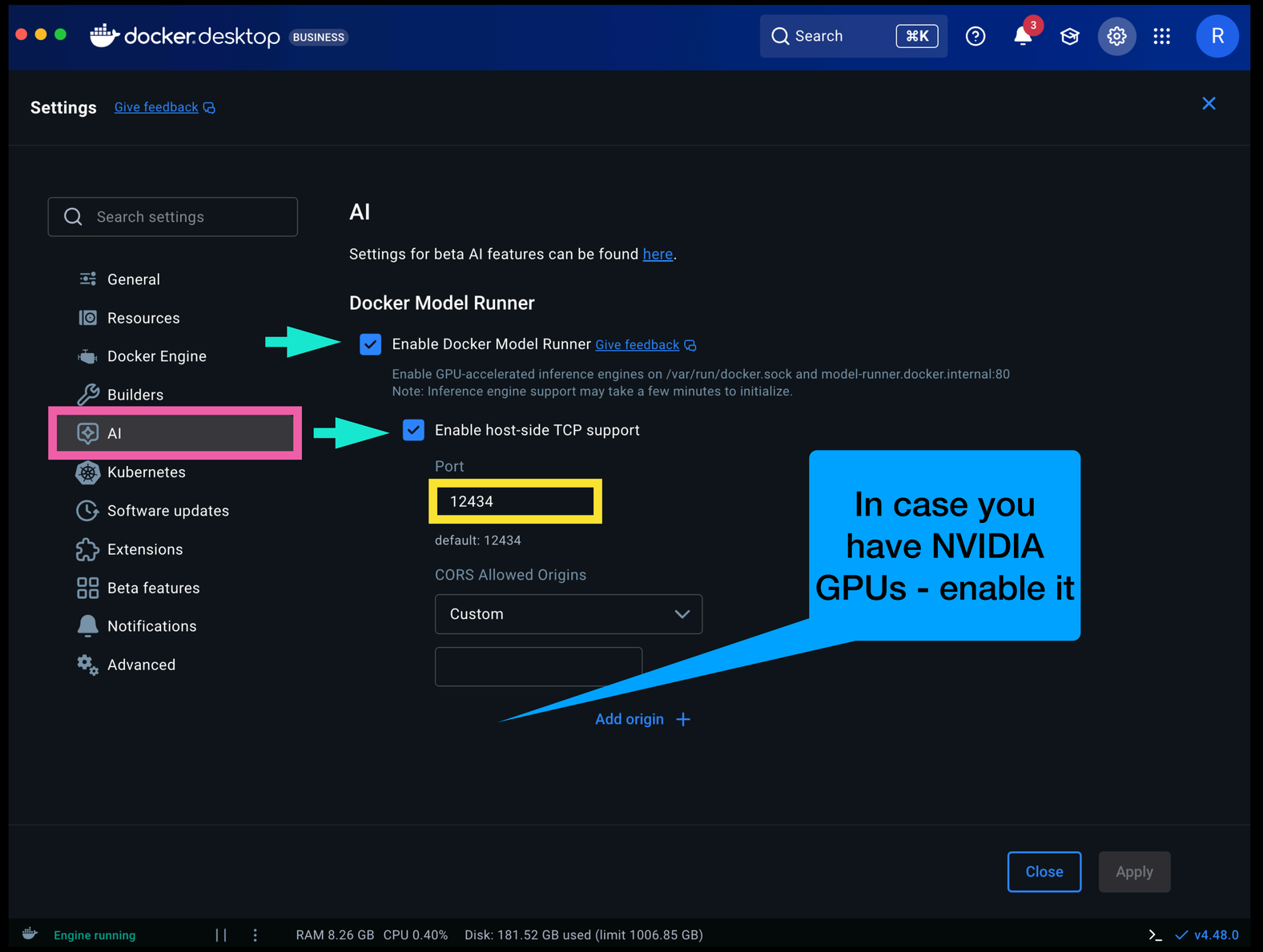
```
paru -S opencode-bin
```

ODSC AI WEST 2025

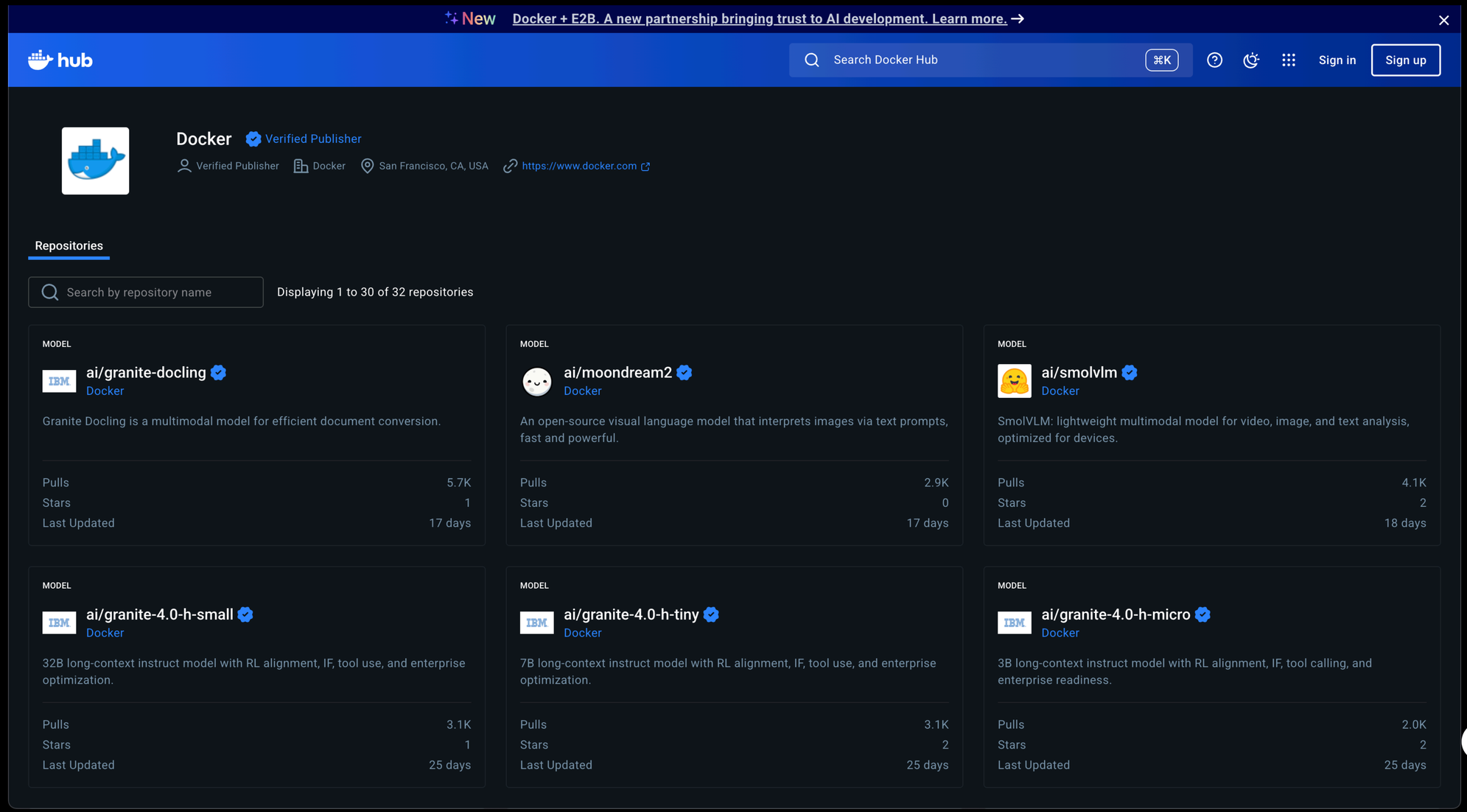
Setup

```
~

ramikrispin ~ ♡17:06 docker login
Authenticating with existing credentials... [Username: rkrispin]

i Info → To login with a different account, run 'docker logout' followed by 'docker login'


Login Succeeded

ramikrispin ~ ♡17:07 hf auth login --token $HF_TOKEN
The token has not been saved to the git credentials helper. Pass `add_to_git_credential=True` in this
function directly or `--add-to-git-credential` if using via `hf`CLI if you want to set the git credent
ial as well.
Token is valid (permission: fineGrained).
The token `DMR Workshop` has been saved to /Users/ramikrispin/.cache/huggingface/stored_tokens
Your token has been saved to /Users/ramikrispin/.cache/huggingface/token
Login successful.
Note: Environment variable`HF_TOKEN` is set and is the current active token independently from the tok
en you've just configured.
```

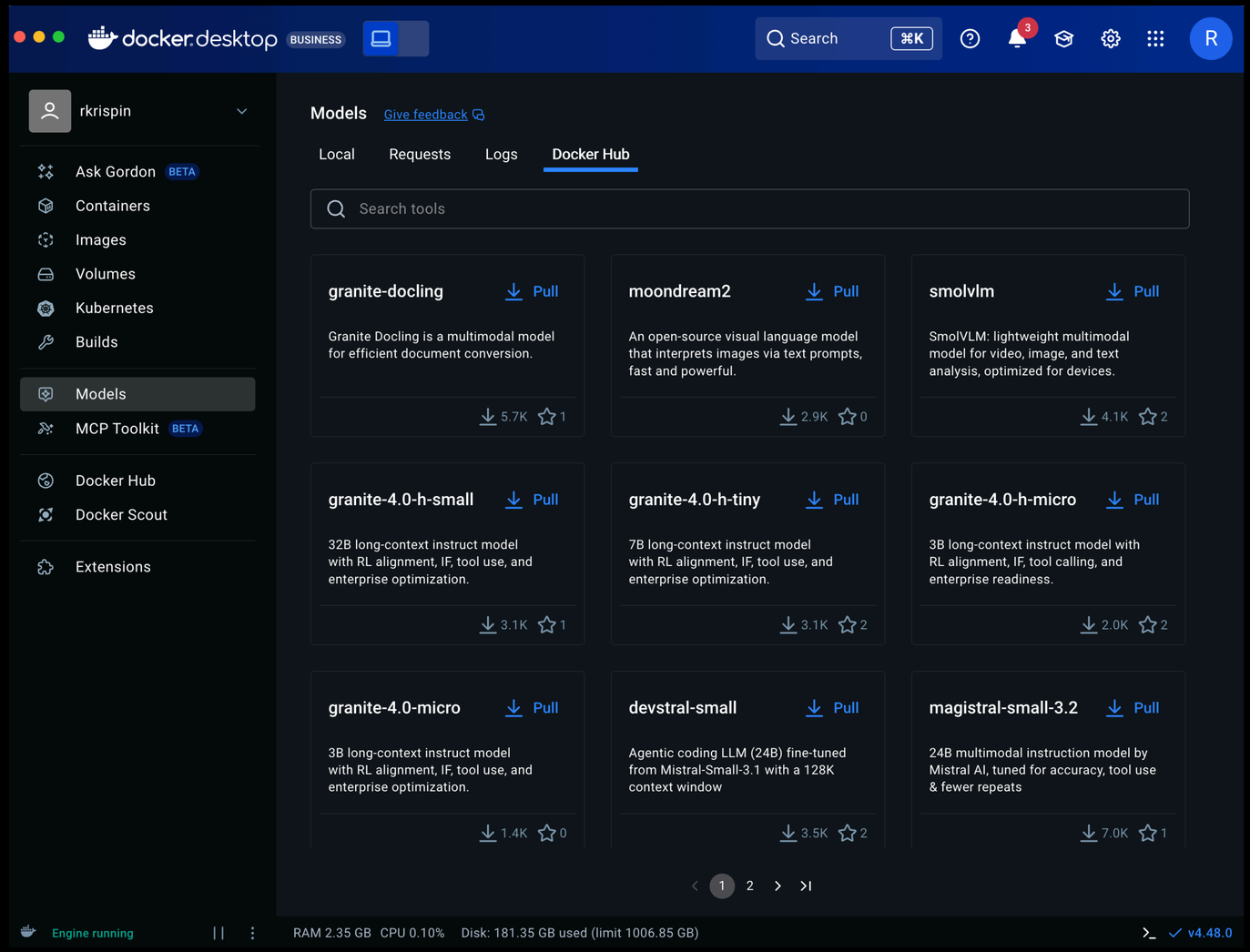
Setup



Models

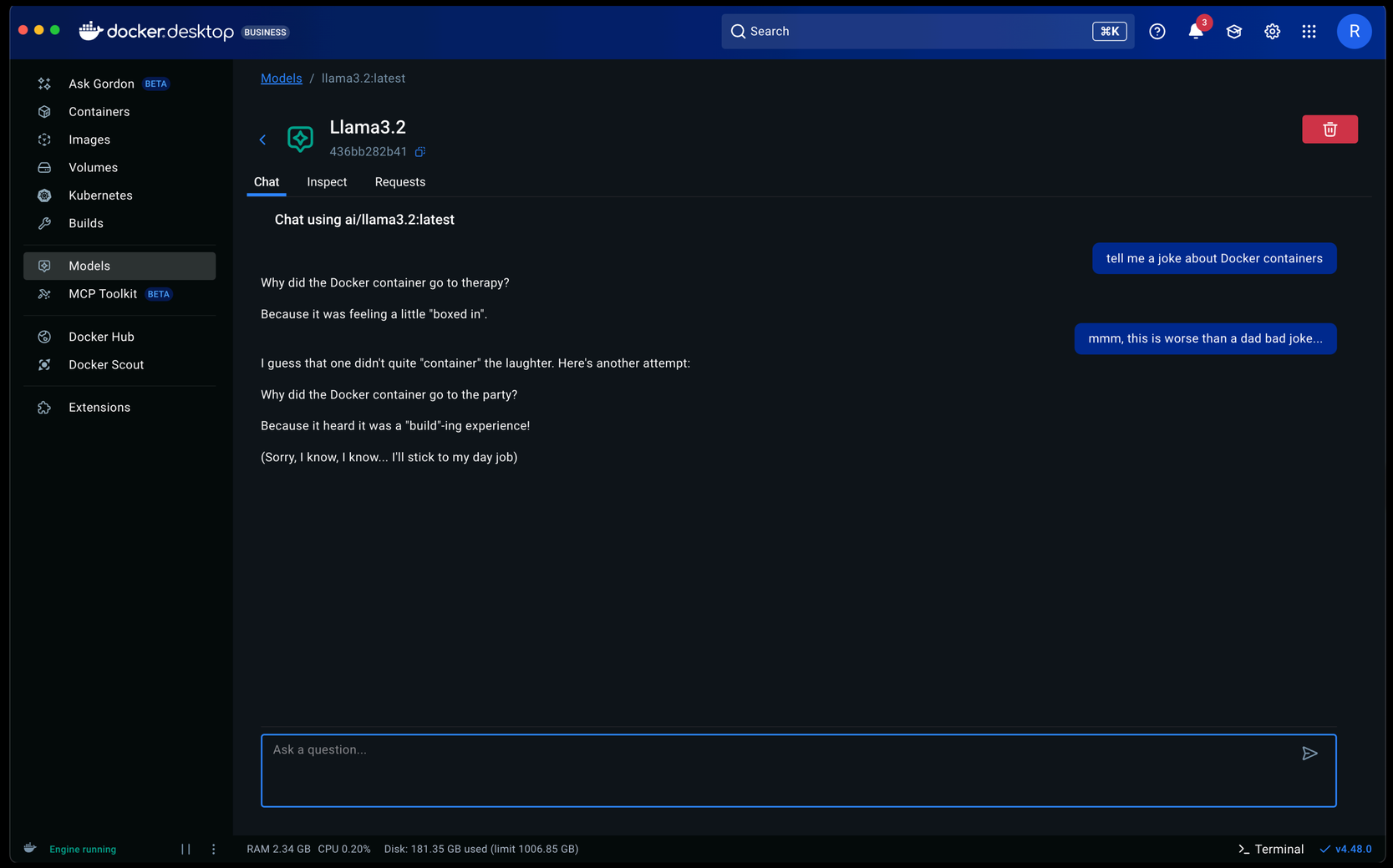


Models



Hello World!

Running LLMs via Docker Desktop



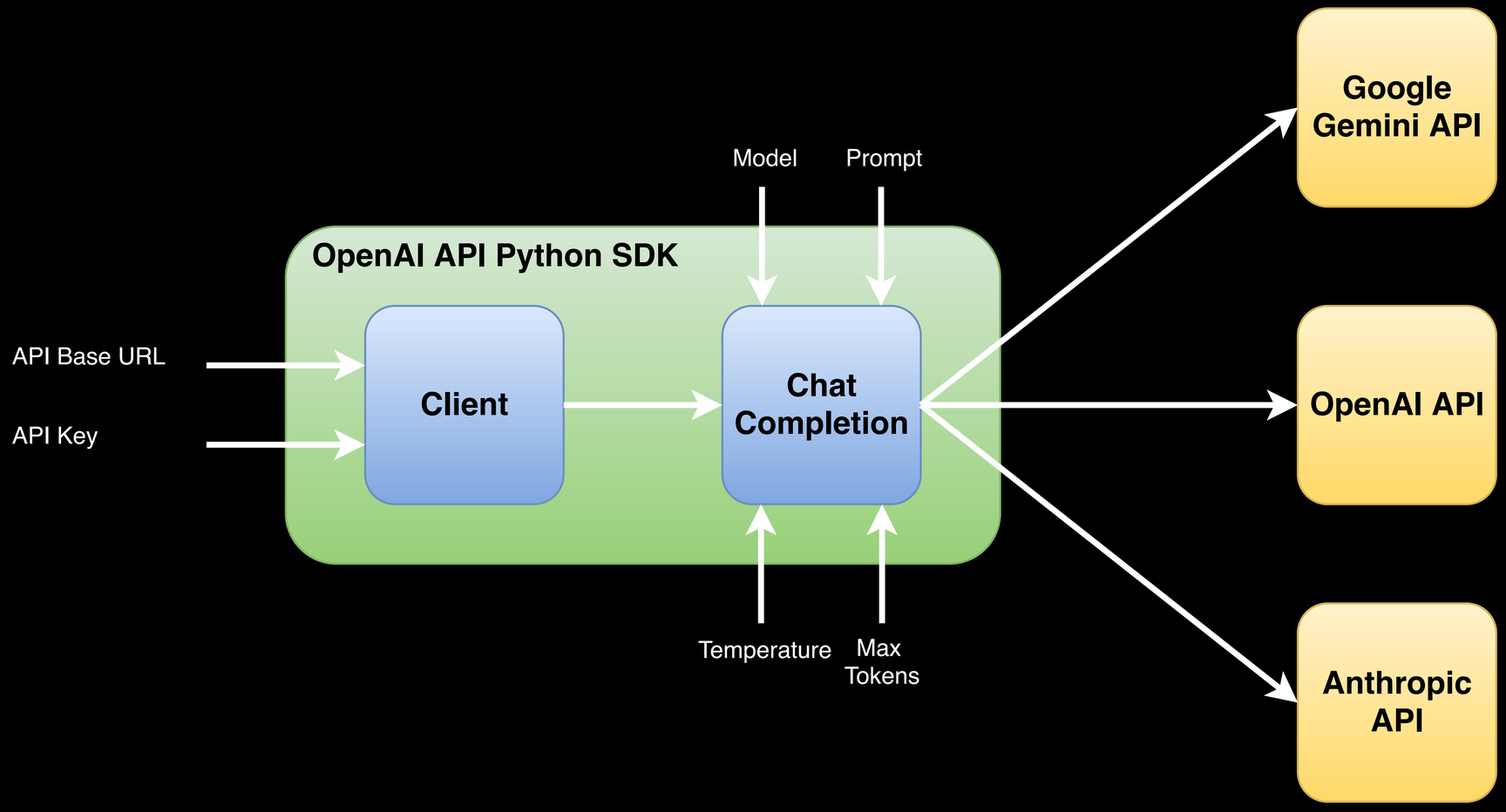
Running LLMs via the CLI

- CLI core commands
- Running LLMs
- Pull models from Docker Hub
- Pull LLMs from Hugging Face
- Update the context window

Running LLMs with Python

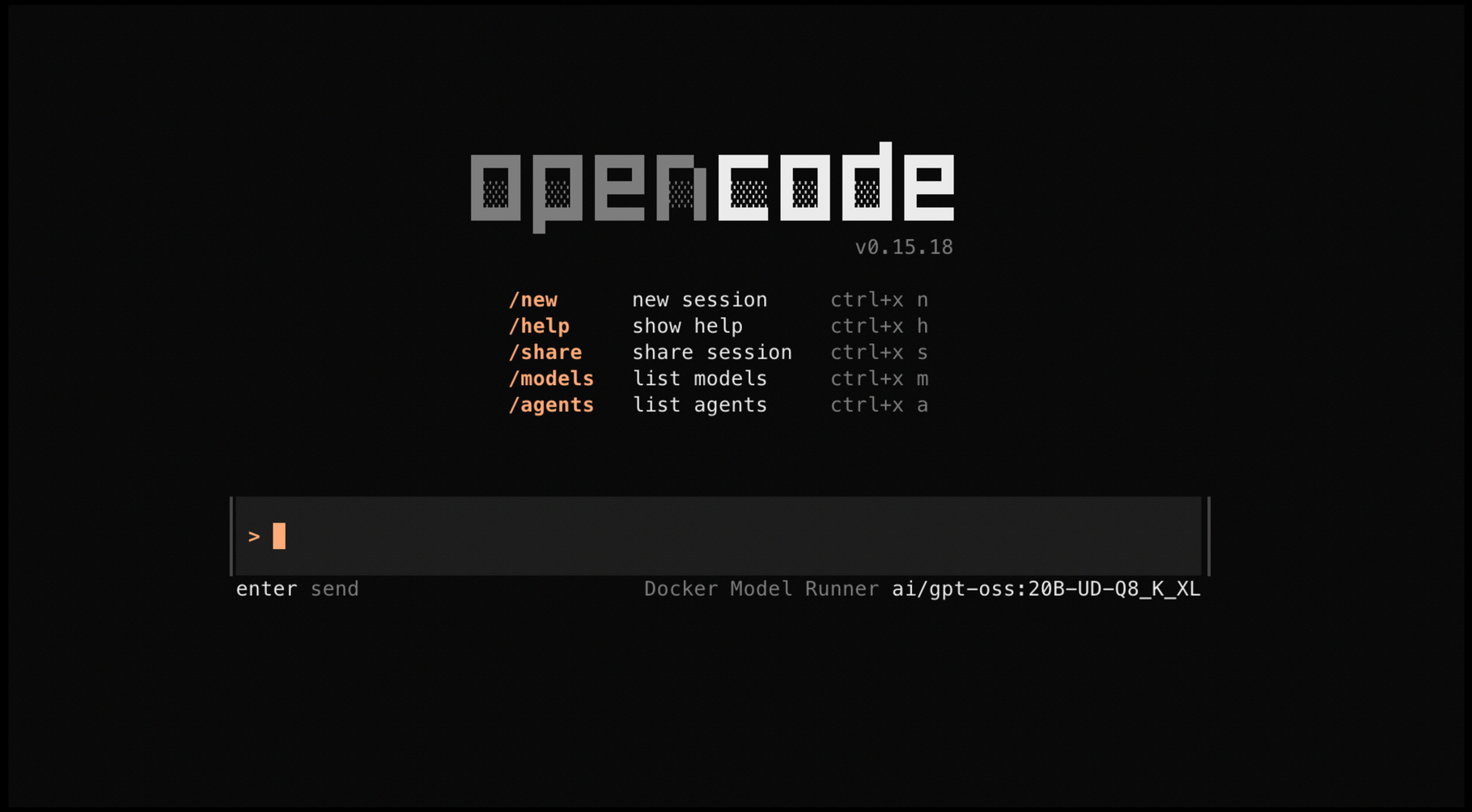
- OpenAI API SDK
- Calling LLMs from Virtual Environment vs. Container
- LangChain

OpenAI API SDK



Demo

OpenCode



OpenCode

```
{
  "$schema": "https://opencode.ai/config.json",
  "provider": {
    "dmr": {
      "npm": "@ai-sdk/openai-compatible",
      "name": "Docker Model Runner",
      "options": {
        // Case using inside a container:
        // "baseURL": "http://model-runner.docker.internal/engines/v1",
        // Otherwise use the following base URL:
        "baseURL": "http://localhost:12434/engines/v1",
        "apiKey": "docker"
      },
      "models": {
        "ai/gpt-oss:20B-UD-Q8_K_XL": {
          "name": "ai/gpt-oss:20B-UD-Q8_K_XL"
        },
        "ai/llama3.2:3B-Q4_0": {
          "name": "ai/llama3.2:3B-Q4_0"
        },
        "ai/devstral-small:24B": {
          "name": "ai/devstral-small:24B"
        }
      }
    }
  }
}
```

Resources

- DMR documentation - https://docs.docker.com/ai/model-runner/
- Docker podcast - https://www.youtube.com/watch?v=9XryO1Oxb4A
- Workshop repo -
- The AIOps Newsletter - https://theaiops.substack.com/